QTL Linkage and Association Analysis
连续性状位点的连锁与关联分析

Jing Hua Zhao (赵京华), Pak Chung Sham (沈柏松)

Section of Genetic Epidemiology/Biostatistics

KCL Insititute of Psychiatry

Aim
目标

- To deliver a brief introduction (简单介绍)

- To bring about problems and discussions (引出问题并启发讨论)

- To build up frequent interactions in the future (建立更多的联系), e-mail: j.zhao@iop.kcl.ac.uk

Structure
结构

- Basic concepts (基本概念)
- Regression, variance component model, quantitative TDT (Haseman-Elston 回归, 方差分量模型以及 TDT)
- Some examples (用例)
- Conclusion (结论)

Basic Concepts
基本概念

- Definition (定义)
- History (历史)
- Major gene (主基因)
- Two-locus, oligogenic, polygenic effect (双位点，若干基因与多基因效应)
- Residual effect (其它效应)

- Hardy-Weinberg Equilibrium/Disequilibrium (H-W 平衡定律 / 不平衡)
- Linkage equilibrium (连锁平衡)
- Genetic heterogeneity (遗传异质性)
- Phenocopy (表现型拷贝)
- IBD/IBS (等位基因的传递 / 状态一致)
- Significance (显著性)
- Power (把握度 / 功效 / 势函数)

Sibling Design and IBD/IBS (同胞对设计及 IBD/IBS)

$12 \rightarrow \leftarrow 34$

$$13 \swarrow \searrow$$

$$\square \quad \begin{cases} 13 & \text{IBD=2} \\ 14 & \text{IBD=1} \\ 23 & \text{IBD=1} \\ 24 & \text{IBD=0} \end{cases}$$

Possible IBD configurations for sib-pairs

Linkage Equilibrium/Disequilibrium (连锁平衡 / 不平衡)

Locus 1 (基因频率): p_1, p_2

Locus 2 (基因频率): q_1, q_2

Disequilibrium parameter (关联参数): d

	q_1	q_2	
	$p_1q_1 + d$	$p_1q_2 - d$	p_1
	$p_2q_1 - d$	$p_2q_2 + d$	p_2
			1

$-\min(p_1q_1, p_2q_2) \le d \le \min(p_1q_2, p_2q_1)$

Study Design
研究设计
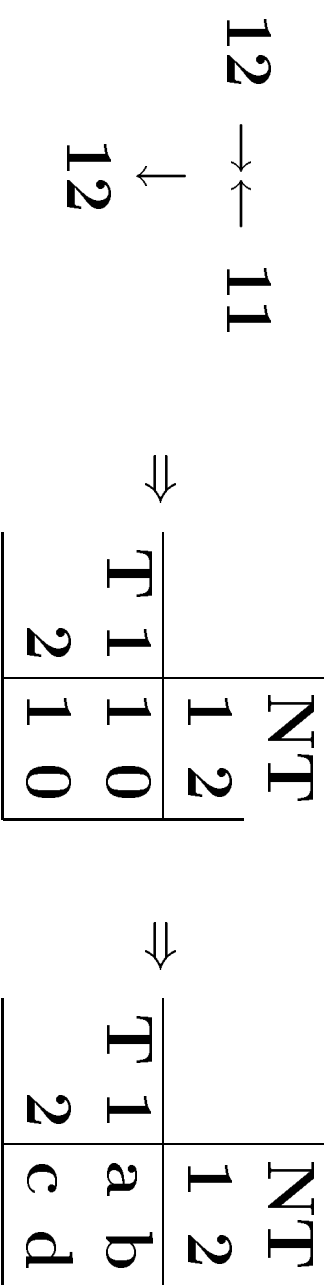
- Population sample (人群样本)
- Siblings (同胞对)
- Family trios (个体及其父母)
- Extended families (大家系)

Typical TDT Design (典型的 TDT 设计)

$12 \not\rightarrow 11$

$12 \downarrow$

$\Rightarrow$

NT	1	2
T 1	1	1
2	1	0

$\Rightarrow$

NT	1	2
T 1	a	b
2	c	d

$\chi_1^2 = \frac{(b-c)^2}{b+c}$, t-test

General Genetic Model
一般遗传模型

$$X = g + G + c + e$$

The phenotype X is additive effects of major gene (g), poly-gene (G), environment (c) and residual (e). (表现型 X 是主基因, 多基因, 环境以及其它效应的总和)

Haseman-Elston Method
Haseman-Elston 方法

$E(Y_j|\pi_j) = \alpha + \beta\pi_j$, $\alpha = \sigma_e^2 + 2\sigma_g^2$, $\beta = -2\sigma_g^2$.

$E(Y_j|\hat{\pi}_j) = \alpha + \beta\hat{\pi}_j$, $\beta = -2(1 - 2\theta)^2\sigma_g^2$.

- Extension (推广): $E[(y_1 - m)(y_2 - m)] = (1 - 2\theta)^2 V_{A\pi} + \text{residual}.$
  Positive regression coefficient indicates linkage (当回归系数正值时有连锁)

- Software (软件): SAGE/SIBPAL

Variance Component Model
方差分量模型

Multinormal for k-th family (第 k 个家系为多元正态)

$$f(X_k; \mu_k, V_k) = \frac{1}{\sqrt{2\pi}|V_k|} e^{-\frac{1}{2}(X_k - \mu_k)'V_k^{-1}(X_k - \mu_k)}$$

Variance-covariance matrix (方差协方差阵)

$$V_k = \hat{\Pi}\sigma_g^2 + 2\Phi\sigma_G^2 + I_{m_k}\sigma_e^2$$

$\hat{\pi}_{ij}$ is estimated proportion of IBD at QTL by individuals i and j, ϕ_{ij} is the kinship coefficient ($\hat{\pi}_{ij}$ 为亲属对在 QTL 的 IBD 比例，ϕ_{ij} 为近缘系数)

Variance-Covariance Matrix for Siblings

同胞对的方差 — 协方差阵

$$V_k = \begin{pmatrix} \sigma_g^2 + \sigma_c^2 + \sigma_e^2 & \pi\sigma_g^2 + \sigma_c^2 \\ \pi\sigma_g^2 + \sigma_c^2 & \sigma_g^2 + \sigma_c^2 + \sigma_e^2 \end{pmatrix}$$

covariance = $[0.5 + (1-2\theta)^2(\pi_m - 0.5)]\sigma_g^2 + \sigma_G^2$ for a linked marker locus m with a recombination rate θ and in linkage equilibrium (标记与性状位点重组率 θ), π_m = the proportions of alleles IBD at marker locus (标记位点的 IBD 比例).

Test of Linkage
连锁的检验

Likelihood ratio statistic is（似然比统计量为）：

$$\lambda = -2(l_0 - l_1)$$

where l_0 and l_1 are the loglikelihoods for null hypothesis $H_0 : \sigma_g^2 = 0$ alternative $H_1 : \sigma_g^2 > 0$（其中 l_0 及 l_1 分别为原假设与备择假设下的对数似然比）

λ approximately 50:50 χ_1^2 and 0, 99% percentile 5.21（λ 近似为 50:50 χ_1^2 与 0, 99% 界值为 5.21）。

Data Transformation and Robustness

数据转换及稳健性

With non-normality, Box-Cox transformation defines as
（非正态数据的转换公式）

$$y = \begin{cases} \dfrac{x^\lambda - 1}{\lambda x_G^{\lambda-1}}, & \lambda \neq 0 \\[2mm] x_G \ln x, & \lambda = 0 \end{cases}, \quad \text{其中 } x_G = \left(\prod_{i=1}^{N} x_i\right)^{1/N}$$

e.g. SAGE/REGC, Stata.

For proper chosen s define （对给定的 s 定义）

$$\delta(r_{jk}) = \begin{cases} -s, & r_{kj} < -s \\ r_{kj}, & -s \leq r_{kj} \leq s \\ s, & r_{kj} > s \end{cases}$$

Quantitative Association Analysis
定量数据的关联分析

Null hypothesis（无效假设）

- Linkage analysis（连锁分析）：$\theta = 0.5$

- Association analysis（关联分析）：$(1 - 2\theta)\delta = 0$
  i.e., association in the presence of linkage, but now focus
  on mean difference rather than frequencies of preferential
  transmission（连锁下检验关联，均值的差异）

Discrete/Quantitative Test – Case–Control Design
(病例对照设计的定性 / 定量检验)

As in the following table, for thresholds l and u (见下表，设定正常或异常标准)

	Discrete(定性)		Quantitative(定量)	
	case(病例)	control(对照)	mean(均值)	variance(方差)
A_1A_1	p^2/K_l	p^2/K_u	$o+a$	s_1^2
A_1A_2	$2pp/K_l$	$2pq/K_u$	$o+d$	s_2^2
A_2A_2	q^2/K_l	q^2/K_u	$o-a$	s_3^2

$$K_l = p^2 l_1 + 2pq l_2 + q^2 l_3, \quad K_u = p^2 + 2pq u_2 + q^2 u_3$$

Discrete/Quantitative Test – Family Trios

(定性／定量检验 – 个体及其父母)

Test genotype relative risks conditional on parental genotype, incorporate population stratification in quantitative test (基于父母的基因型构造似然比统计量，从而检验基因型之间相对危险度的大小；定量的检验还可以考察人群的分层／混合效应).

Schaid and Sommer (1993), Knapp (1995), van den Oord (1999)

Multi-allelic TDT
多个等位基因位点的 TDT

Resemblance to Spielman and Ewens (1996) and two-group t-test (Xiong et al. 1998; Wang and Cohn 1999)

$$TDT_m = \frac{m-1}{m} \sum_{i=1}^{m} \frac{(\bar{Y}_{i.} - \bar{Y}_i)^2}{S_i^2(n_{i.}^{-1} + n_{.i}^{-1})}$$

Let n_{ji} **heterozygous parents transmit allele** M_j, k**th child's trait value** Y_{jik}, $n_{i.} = \sum\limits_{\substack{j=1 \\ j \neq i}}^{m} n_{ij}$, $n_{.i} = \sum\limits_{\substack{j=1 \\ j \neq i}}^{m} n_{ji}$, $\bar{Y}_{i.} = n_{i.}^{-1}\sum\limits_{\substack{j=1 \\ j \neq i}}^{m}\sum\limits_{k=1}^{n_{ij}} Y_{ijk}$, $\bar{Y}_j = n_{.i}^{-1}\sum\limits_{\substack{j=1 \\ j \neq i}}^{m}\sum\limits_{k=1}^{n_{ji}} Y_{jik}$, $S_i^2 = (n_{i.} + n_{.i} - 2)^{-1}\left[\sum\limits_{\substack{j=1 \\ j \neq i}}^{m}\sum\limits_{k=1}^{n_{ij}}(Y_{ijk} - \bar{Y}_{i.})^2 + \sum\limits_{\substack{j=1 \\ j \neq i}}^{m}\sum\limits_{k=1}^{n_{ji}}(Y_{jik} - \bar{Y}_{.i})^2\right]$

A Summary of Computer Programs

常用计算机程序

- ACT – ML/ML variance component model by Amos (1994)

- FISHER – variance component model by Lange

- GENEHUNTER – Haseman-Elston method and variance component model

- Loki – Markov Chain Monte Carlo (MCMC) linkage

- Mapmaker/SIBS – Haseman-Elston method

- Mx/MxGUI – structual equation modelling (SEM)

- NOCOM – EM algorithm with unrelated individuals
- PAP – ordinary linkage and multifactorial trait
- PATHMIX – path analysis of nuclear families
- POINTER – segregation analysis
- QTDT – a collection of quantitative TDT programs
- SAGE – Modular, handle covariates
- SOLAR – sequential oligogenic linkage analysis routines
- XDT – discrete TDT/SDT and Rabinowitz test

Applications
应用实例

- Twin EPQ data (Sham 1998)
- Collaborative Study on the Genetics of Alcoholism (COGA)
- Genetic-Environmental Nature of Emotional States in Siblings (GENESIS)
- ESF project on Schizophrenia/Bipolar/Psychosis
- Angiotensinogen among Nigerians

Twin EPQ Data Analysis
双生子 EPQ 资料分析

- 522 female MZ, 272 female DZ twin-pairs
- Detailed from Sham (1998) Chapter 5.
- Wide range of biometric models
- SPSS, SAS programs and Mx scripts
  http://alpha.iop.kcl.ac.uk/jinghua
  http://chsi.moh.gov.cn/jhz0

COGA data
COGA 资料分析

- Genetic Analysis Workshop 11 (GAW11)
- Extended families, DSMIIIR and Feighner criteria
- ERP as QTL, genome scans
- Power analysis by Williams et al. (2000), discrete trait by Curtis et al. (2000)

GENESIS

- Aim (目标) – identify QTLs
- Design (设计) – siblings
- Materials (材料)
- Methods (方法) – SEM, linkage and association
- Result (结果)

ESF Project

- Data (资料) – 137 families, 388 markers

- Methods (方法)
  - linkage (**LINKAGE/FASTLINK, GENEHUNTER**)
  - association (**ASPEX**)

- **Result (结果) – 10 hotspots MOD>2, biggest 4.16; 1,000 simulations by SLINK, 52 lod in [1,2), 5 in [2,3), 0 equal or greater than 3; 3-point MOD 4.44 for psychosis, heterogeneity test nonsignificant**

Angiotensinogen and Hypertension
血管紧张素与高血压

- The Framingham Study

- Chinese sample (M135, T174) – Affected pedigree method (APM)

- Nigerian sample (AHG 1999,63:293-300) – SAGE/REGC

Conclusion – Limiting Factors
结论（影响因素）

- Power and sample size（设计的功效及样本量）, analytically/by simulation（解析求解或随机模拟）.

- Feasibility of mathematics, computing（数学，计算机的可行性）. Large sample for linkage analysis for small effects whereas dense map for association methods（大样本检测微小效应或密集的基因图图）.

- Population genetics（群体遗传学）

References

文献

- Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics. Longman.

- Gillespie J (1998) Population Genetics. The Johns Hopkins University Press

- Khoury MJ, Beaty TH, Cohn BH (1993) Fundamentals of Genetic Epidemiology. The Johns Hopkins University Press

- Kempthorne O (1957) An Introduction to Genetic Statis-

tics. Wiley.

- Lange K (1997) Mathentical and Statistical Methods for Genetic Analysis. Springer

- Ott J (1998) Analysis of Human Genetic Linkage. The Johns Hopkins University Press

- Sham P (1998) Statistics in Human Genetics. Arnold

- Weiss K (1993) Genetic Variation and Human Disease. Cambridge University Press

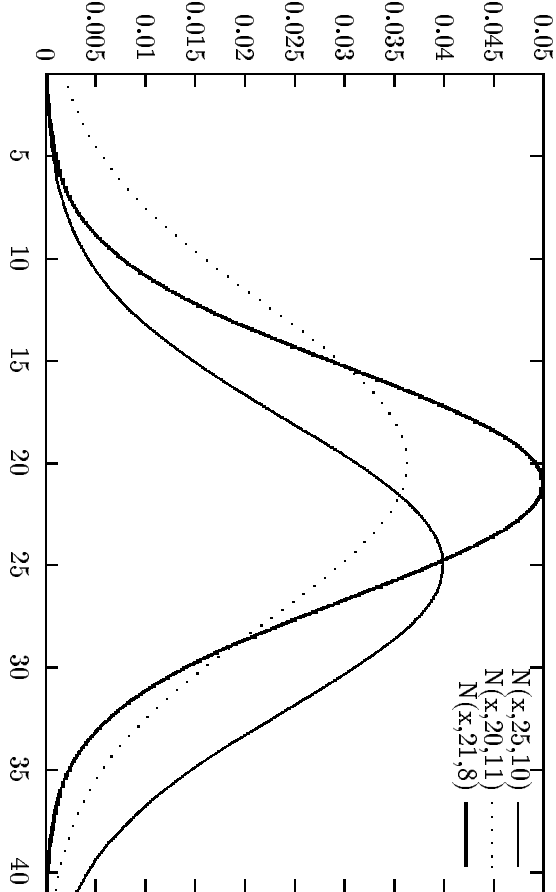- Weir, BS(1996) Genetic Data Analysis II. Sinauer Associates.

URLs

Internet 地址

- **Online Mendelian Inheritance in Man (OMIM)**
  http://www3.ncbi.nlm.nih.gov/omim (JHU)
  http://www.hgmp.mrc.ac.uk/omim (UK)

- **List of linkage programs**
  http://linkage.rockefeller.edu

- **Online journals**
  www.iop.kcl.ac.uk/IoP/Departments/PsychMed/GepiBSt/index.stm

Normal Distributions (正态分布)

Distribution of liability
（阈值 T 下人群中易患性的分布）